DOCUMENT RESUME

ED 169 870 HE 011 269

AUTHOR Chase, Clinton I.; Keene, John M., Jr.
TITLE Validity of Student Ratings of Faculty. Indiana Studies in Higher Education, Number Forty.
INSTITUTION Indiana Univ., Bloomington. Bureau of Educational Studies and Testing.
PUB DATE 79
NOTE 14p.
AVAILABLE FROM Bureau of Evaluative Studies and Testing, Division of Research and Development, Indiana University, Bloomington, Indiana 47401

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Correlation; Evaluation Methods; Faculty Evaluation; *Failure Factors; Higher Education; Institutional Research; Questionnaires; Rating Scales; Research Projects; Student Characteristics; Student Evaluation; *Student Evaluation of Teacher Performance; *Student Teacher Relationship; *Success Factors; *Validity

ABSTRACT

A study using two types of analyses to examine the validity of student ratings of faculty is described. The first analysis examined faculty ratings in terms of how high-achieving sections of courses rate their instructors in contrast with how low-achieving sections rate their instructors to determine if those teachers who advance their students the farthest receive higher ratings than those teachers whose classes do not advance as far. The results on the total scores of the rating scales showed high-achieving sections of courses reporting stronger ratings for their instructors than was true of low-achieving sections. The second analysis identified high achievers as those students who did better than they were predicted to do, and low achievers were those students who did worse than they were predicted to do. This analysis focused on individual student achievement. On individual items and on total scale scores high achievers provided more substantial ratings than did low achievers. It is suggested that if the circumstances in a class lead a student into achieving better than is predicted, that student may rate the instructor higher than the student who achieves below expectation. It is suggested that the validity of the evaluation procedure is demonstrated through these analyses. (BH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED169870

# VALIDITY OF STUDENT RATINGS OF FACULTY

CLINTON I. CHASE
JOHN M. KEENE, JR.

INDIANA STUDIES IN HIGHER EDUCATION
NUMBER FORTY

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ind. University

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

HE 011269

BUREAU OF EVALUATIVE STUDIES AND TESTING
DIVISION OF RESEARCH AND DEVELOPMENT
INDIANA UNIVERSITY
BLOOMINGTON, INDIANA
1979

TABLE OF CONTENTS


Introduction . . . . . 1

Procedure . . . . . 2

Results

First Analysis . . . . . 3

Second Analysis . . . . . 6

Discussion . . . . . 10

Conclusions . . . . . 12

References . . . . . 12


LIST OF TABLES


| Table | | Page |
|---|---|---|
| 1 | An analysis of differences in faculty rating total scores among a group of "high" achieving sections and "low" achieving section in German and in accounting . . . . . | 3 |
| 2 | Means of high and low achieving groups by course . . . . . | 4 |
| 3 | Means and standard deviations for items by course and by high and low achieving groups . | 5 |
| 4 | Analysis of variance testing differences in faculty ratings among course areas (accounting, German, mathematics, and sociology) and between achievers and non-achievers . . . . . | 7 |
| 5 | Mean faculty ratings provided by achievement groups and by course groups . . . . . | 7 |
| 6 | Item ratings by high and low achievers within each course . . . . . | 9 |




2-A

## Introduction

Teacher and course evaluation systems, in which students rate various aspects of a faculty member's performance, are in wide use. Although a number of studies have been done on such ratings (reviews provided by McKeachie, 1969; Costin, Greenough, and Menges, 1971; and Menges, 1973), validity of ratings is still a matter of controversy. Rodin and Rodin (1972) provided often quoted data that correlated negatively with teaching quality. These findings have been contested by Rodin, Frey, and Gessner (1975), while further studies (Frey 1973; Gessner 1973; Sullivan and Skanes, 1974) did in fact produce positive relationships between ratings and teaching ability. Frey, Leonard & Beatty (1975) with data collected at three universities, concluded that core factors in rating scales did indeed correlate positively with teacher effectiveness. On the other hand, Turner & Thompson's (1974) findings were supported by Greenwood, Hazelton, Smith & Williams (1976) in showing a negative relationship.

The above research indicates the lack of concensus as to the validity of student ratings of instructor and course effectiveness. However, some of the conflicting data may relate to the type of course involved. The nature of effective teaching in one subject area may deviate widely from effective teaching in another area. For example, the instructor behavior that brings about greatest learning in literature may not be the behavior that fosters learning in chemistry. Further, some of the above discrepancies in results may also be tied to experimental design factors that define teaching effectiveness. For example, grouping students into "high" and "low" final test scores may in fact be grouping on academic aptitude, not on amount of learning. This points to differing methods of group selection as an explanation of discrepant findings. If one study uses raw gains as the basis for selecting high and low achieving groups, the distribution of faculty ratings may indeed be different from those in a second investigation where adjusted gains were used, or where a regression model was the basis for identifying "high" and "low" groups.

The following study was designed to further assess the validity of student ratings of faculty. A set of faculty rating statements were asked to discriminate between high and low achievers as defined by each of two criteria: 1) sections of courses were sorted into high and low achieving classes based on the mean score on the common final test, adjusted for talent factors; and 2) high and low achievers were identified as those individual students whose final test score was higher or lower than predicted from talent factors. The first analysis was called high and low *class* achievement; the second high and low *individual* achievement. In each analysis faculty ratings were observed to see if they did indeed show differences between "high" and "low" groups of achievers. The study also looked at ratings across academic courses to see whether or not student ratings did indeed vary with subject matter.

## Procedure

Four multi-section courses were involved in the study: accounting with 16 sections, German with 8 sections, mathematics with 4 sections, and sociology with 4 sections. SAT scores, high school ranks, and pretest and posttest scores in each course were collected for students in all classes. Also, a common eleven item faculty rating scale was applied across these courses. Since certain data were missing for some students in all classes, and since in certain classes data were missing on one or more entire variables, the analysis did not include all students or all sections.

The first step in the analysis based on class achievement was to identify extreme achievement classes after posttest scores were adjusted for potential. Regression procedures were used to adjust posttest scores for portions of posttest score variance which could be accounted for by weighted combinations of SAT scores, high school rank, and pretest data. Then, within a course the faculty ratings of high achieving sections were compared with faculty ratings of the low achieving sections.

Full data were available for six accounting sections, eight German, four mathematics and four sociology sections. To maximize mean achievement differences only the highest and lowest achievement sections were used. Since mathematics and sociology had only four sections each, this left only one "high" and one "low" achieving section for each course. However, the middle sections could be dropped in accounting, leaving two "high" and two "low," and in German leaving three "high" and three "low." Therefore, because of the larger numbers of observations possible, the analysis of faculty ratings in this part of the study was based on data collected from students in German and in accounting classes, only.

The second phase of the study applied the individual achievement criterion for "high" and "low" groupings. A procedure cited by Cronbach and Furby (1970) was employed at this point. The pretest scores in the separate courses, SAT scores, and ranks in the high school classes were used to predict a final test score for each student in each course. Then the predicted final test score was subtracted from the student's actual final test score. (A separate prediction equation was developed for each of the four courses.) To insure that group achievement differences were more than casual, students whose final test scores were near their predicted score (within ±.33 of a standard error) were dropped out. This left two groups: those students who performed clearly above their expected level ("high" achievers), and those who performed clearly below their expected level ("low" achievers).

In both above designs, a two dimensional analysis of variance (achievement level by course group) was performed on a scale composed of the eleven faculty rating items designated by the College of Arts and Sciences as the core

for all A&S faculty. Instructors in the study could choose additional rating items if desired, but all instructors used the same basic core of eleven items.

The intent of this study was to validate faculty ratings. If ratings are indeed valid they should show differences between those submitted by students in "high" achieving sections in the course and those submitted by students who were in "low" achieving sections. Also, items should discriminate between those students who, on their final test, scored above and those who scored below their predicted level.

The first analysis resulted in a two-by-two analysis of variance design (high-low achievers and German-accounting courses), with the analysis being done on the faculty ratings. The second analysis was a two-by-four analysis on scores on the faculty rating items, with two student achievement groups (above expected, below expected) as one variable of classification, and the four courses (German, accounting, mathematics, sociology) as the second variable of classification.

## Results

### The first analysis.

Eight German and six accounting sections has sufficiently complete data for this analysis. The two classes out of eight in German with the highest adjusted final test mean were called "highs" and the two with the lowest adjusted final test mean were called "lows." The six sections of accounting were similarly divided, and there again the two sections with highest adjusted final means were called "highs" and the two lowest scoring sections were called "lows." Since all students had responded to a common set of eleven faculty rating items, the analysis was based on a comparison of the faculty ratings given by "highs" and by "lows" across both courses. The results are in Table 1.

Table 1

An analysis of differences in faculty rating total scores among a group of "high" achieving sections and "low" achieving sections in German and in accounting.

| Source | ss | df | ms | F | Signif. |
|---|---|---|---|---|---|
| between achiev. gps | 227.603 | 1 | 227.603 | 11.030 | .002 |
| between courses | 14.219 | 1 | 14.219 | .689 | n.s. |
| interaction achiev. by course | 147.300 | 1 | 147.300 | 7.139 | .01 |
| residual | 1258.680 | 61 | 20.634 | | |

These data indicate: 1) that high achieving sections produced a different level of ratings than did low achieving sections, and 2) that the level of ratings within achievement groups was different across courses. These findings are elaborated in Table 2, where group means on the faculty rating scale are shown for high and low achievers in each of the two courses.

The difference in faculty ratings given by high and low achieving classes in German is somewhat larger than the difference between high and low achievement groups in accounting (see Table 2). Also, the low sections in accounting rated their instructors higher than did the lows in German, but highs in German rated their instructors above the accounting highs. This is reflected in the significant interaction reported in Table 1. It appears that the magnitude of the difference in ratings given by high achieving classes over low achieving classes may be tied to the course in which the ratings are given.

Table 2

Means of high and low achieving groups by course.

| Course | High | Low |
|---|---|---|
| German | 48.25 | 39.73 |
| Accounting | 45.48 | 43.64 |

The above analysis is based on a *total* rating score summed across eleven standard rating items. A second analysis was also made; this time each of the eleven items were analyzed separately. It is recognized that the reliabilities of individual items is somewhat lower than for the total score, so the results of this second analysis will be used primarily as a basis for hypothesis.

Significant differences (.05 level) were found between high and low achieving groups on the following items:

2. I knew what was expected of me in this course.
4. My instructor seems enthusiastic about teaching this course.
5. My instructor is well prepared for class meetings.
6. My instructor is able to explain the subject clearly.
7. My instructor makes the subject interesting.
11. My instructor ranks among the best at this level.

Mean item scores - rated from 1 (low) to 5 (high) are given for each of the above items in Table 3. Here the data show that the high achieving sections consistently rated their instructors more positively than did the low achieving sections on six of the eleven items, and that the differences

between achievement groups are typically greater for students in German than in accounting. This disparity between ratings in the two courses is the reason a significant interaction was found between course and achievement groups in items 4, 5, 6, 7, and 11.

Table 3

Means and standard deviations for items by course and by high and low achieving groups.

| Item No. | Course | High Achieving Mean | S.D. | Low Achieving Mean | S.D. |
|---|---|---|---|---|---|
| 2 | German | 4.38 | .74 | 3.77 | 1.01 |
| | Accounting | 4.38 | .67 | 3.80 | .87 |
| 4 | German | 4.75 | .46 | 3.55 | .82 |
| | Accounting | 4.38 | .74 | 4.36 | .49 |
| 5 | German | 4.75 | .46 | 3.73 | .79 |
| | Accounting | 4.76 | .44 | 4.60 | .50 |
| 6 | German | 4.63 | .52 | 3.45 | .93 |
| | Accounting | 4.43 | .68 | 4.32 | .69 |
| 7 | German | 4.63 | .52 | 3.18 | .98 |
| | Accounting | 4.38 | .80 | 3.92 | .64 |
| 11 | German | 4.75 | .46 | 3.09 | .83 |
| | Accounting | 4.38 | .74 | 4.08 | .81 |

An analysis of the items that discriminated between high and low achieving groups appears to center on items that reflect classroom delivery procedures. The class knew where it was going, the instructor was prepared, explained topics well, was enthusiastic about his work, and made the class interesting. This seems to suggest that instructors whose sections achieved highly concentrated on ordering the content into deliverable units which are predictable by the students, and delivering these units with a sense of conviction.

The rating items that did not produce significance between achievement groups were:

1. I learned a lot in this course.
3. This course is one of the most difficult at this level.
8. The exams covered the most important aspects of the course.
9. Feedback on exams indicates my standing in the course.
10. My instructor is regularly available for consultation.

These items indicate that high and low achieving groups did not rate differently the evaluative activities of an instructor (items 8, 9, and possibly 1). Nor did the student's perception of the magnitude of the course (items 1 and 3) rate differently between high and low achieving groups.

The item concerning the availability of the instructor, which failed also to discriminate, may suggest that because one does not perform well in the classroom does not mean that one will be less effective in outside-of-class contacts. It may also say that students do not see out-of-class contact as a frequently needed event, hence not important.

In summary, the high achievers point to qualities of their instructor's daily classroom performance; however, assessment of the student's work and the instructor's out-of-class accessibility do not seem to be regarded differently by high and low achievers.

The second analysis

This analysis differed from the first in how "high" and "low" achievement groups were defined. In the first analysis reported above, the class achievement criterion was used. High and low sections were then compared on their faculty ratings.

In the second analysis *individual students* (rather than sections) were designated as "high" and "low" achievers, depending on whether their final test scores were substantially higher than predicted, or lower than predicted. Predictions were based on the best combination of SAT-V and M scores, rank in the high school class, and pretest data in the course. Separate prediction equations were made up for the four courses: German, accounting, mathematics, and sociology.

The question pursued by this analysis is: Do students who achieve better than they are predicted to do in a given course rate their instructors higher than students who achieve lower than they are predicted to do? This question assumes that selected qualities in an instructor's style and procedure may bring a substantial response from some students, but may not stimulate full achievement by other students. If this is so, a valid faculty rating scale should show "over achievers" providing a different set of ratings for faculty than will the "under-achievers."

As cited above, to increase the likelihood that the students in the high group were really achieving above expectation, and those in the low group were achieving below expectation, all students whose final test score was near expected levels were eliminated. Analysis of variance procedures were then applied on the faculty rating data for the remaining high and low achieving groups. The eleven rating statements used above were analyzed as a total scale, and (with reservations about reliability) as separate items. This time all four courses - German, accounting, mathematics, and sociology - were involved.

When the data were analyzed for total rating scale scores, a significant difference was found between achievement groups and for courses, but not for interaction between

courses and achievement (Table 4). The high achievers rated their instructors more favorably than did low achievers; students in some courses rated their instructors more favorably than students in other courses. These data are given in Table 5. Instructors in accounting and German received more favorable ratings than those in mathematics and sociology. It should, however, be noted that the number of students in the mathematics analysis was very small, which may dilute the reliability of these findings.

Table 4

Analysis of variance testing differences in faculty ratings among course areas (accounting, German, mathematics, and sociology) and between achievers and non-achievers.

| Source | ss | df | ms | F | Signif. |
|---|---|---|---|---|---|
| between Ger., Acctg. | 715.23 | 3 | 238.41 | 8.01 | .001 |
| between achiev. gps. | 221.58 | 1 | 221.58 | 7.45 | .007 |
| interaction course X achievement | 11.23 | 3 | 3.74 | .13 | .945 |
| residual | 5268.02 | 177 | 29.76 | | |

Table 5

Mean faculty ratings provided by achievement groups and by course groups (maximum 55 points)

| | Means |
|---|---|
| Achievement | |
| Hi | 41.62 |
| Lo | 39.54 |
| Course | |
| Accounting | 42.36 |
| German | 42.41 |
| Mathematics | 34.09 |
| Sociology | 40.42 |

Again items were analyzed separately, but as before the reliability of separate item data is less than for the total scale. Therefore, these results are to be taken with reservation. The following items produced significant differences between high achievers and low achievers under the second procedure described above:

1. I learned a lot in this course.
4. My instructor seems enthusiastic about this course.
5. My instructor is well prepared for class meetings.

7. My instructor makes the subject interesting.
9. Feedback on the exams indicated my standing in the course.
11. My instructor ranks among the best at this level.

Items which showed no differences between achievement levels were:

2. I knew what was expected of me.
3. This course is one of the most difficult at this level.
6. My instructor is able to explain the subject clearly.
8. The exams covered the most important aspects of the course.
10. My instructor is regularly available for consultation.

It appears that high achievers differ from low achievers in their perception of the instructor as being prepared, interesting, and enthusiastic, who also provides feedback. How well the exams cover the course, availability of the instructor outside of class, the perceived difficulty of the course were regarded essentially the same by both achievers and non-achievers.

Several items received different ratings in some courses than in others. These items are marked with a plus (+) in Table 6. The following items showed significant differences in how students rated their instructors across courses:

1. I learned a lot in this course - Accounting and German students tended to rate this item high; sociology and mathematics rated it moderately.
3. This course is one of the most difficult at this level - Accounting and German students ranked this item moderately high, while mathematics and sociology rated it low.
4. My instructor seems enthusiastic about teaching this course - Sociology and German rated their instructors high on this item, with accounting close behind, but mathematics students ranked their instructors low.
5. My instructor is well prepared for class meetings - All courses rated their instructors high except mathematics where the ratings were substantial but not high.
6. My instructor is able to explain the subject clearly - On this item all courses received a moderately high rating except mathematics, which received somewhat lower ratings.
7. My instructor makes the subject interesting - On this item all departments received a moderately high rating except mathematics where ratings averaged somewhat lower.
11. My instructor ranks among the best at this level - German students ranked their instructors highest, mathematics lowest, and accounting and sociology were between the extremes.

It should be pointed out that the above rating analysis is not a universal endorsement for or against any course.

Table 6

Item ratings by high and low achievers within each course.

| Item No. | Course | High Achievers N | Mean | S.D. | Low Achievers N | Mean | S.D. |
|---|---|---|---|---|---|---|---|
| 1*+ | Acctg | 25 | 4.28 | .89 | 18 | 3.94 | .73 |
| | Germ | 18 | 4.11 | .47 | 22 | 4.14 | .60 |
| | Math | 6 | 3.00 | 1.26 | 5 | 3.00 | 1.00 |
| | Soc | 76 | 3.55 | .92 | 38 | 3.05 | .99 |
| 2 | Acctg | 25 | 4.04 | .73 | 18 | 3.78 | 1.22 |
| | Germ | 18 | 3.72 | .89 | 22 | 3.95 | .72 |
| | Math | 6 | 3.67 | 1.03 | 5 | 3.60 | .55 |
| | Soc | 76 | 3.97 | .83 | 38 | 3.79 | .96 |
| 3+ | Acctg | 25 | 3.48 | 1.29 | 18 | 3.94 | .87 |
| | Germ | 18 | 3.72 | .89 | 22 | 3.55 | 1.01 |
| | Math | 6 | 2.33 | 1.03 | 5 | 3.20 | 1.64 |
| | Soc | 76 | 2.61 | .98 | 38 | 2.79 | .98 |
| 4*+ | Acctg | 25 | 3.92 | .86 | 18 | 3.83 | .79 |
| | Germ | 18 | 4.44 | .51 | 22 | 4.18 | .85 |
| | Math | 6 | 2.67 | 1.03 | 5 | 2.40 | .89 |
| | Soc | 76 | 4.29 | .76 | 38 | 4.00 | .84 |
| 5*+ | Acctg | 25 | 4.28 | .68 | 18 | 4.28 | .83 |
| | Germ | 18 | 4.56 | .51 | 22 | 4.55 | .51 |
| | Math | 6 | 3.66 | 1.03 | 5 | 3.20 | 1.30 |
| | Soc | 76 | 4.62 | .56 | 38 | 4.34 | .61 |
| 6+ | Acctg | 25 | 4.24 | .72 | 18 | 3.89 | 1.13 |
| | Germ | 18 | 3.89 | .68 | 22 | 4.18 | .66 |
| | Math | 6 | 3.00 | 1.26 | 5 | 2.40 | 1.14 |
| | Soc | 76 | 4.20 | .86 | 38 | 4.05 | .80 |
| 7*+ | Acctg | 25 | 3.76 | 1.13 | 18 | 3.18 | 1.13 |
| | Germ | 18 | 3.94 | .64 | 22 | 3.77 | .87 |
| | Math | 6 | 2.67 | 1.21 | 5 | 2.20 | .84 |
| | Soc | 76 | 3.75 | 1.00 | 38 | 3.14 | 1.16 |
| 8 | Acctg | 25 | 3.84 | .90 | 18 | 3.72 | 1.07 |
| | Germ | 18 | 3.33 | .91 | 22 | 3.23 | .97 |
| | Math | 6 | 3.33 | 1.03 | 5 | 3.40 | 1.34 |
| | Soc | 76 | 3.39 | .95 | 38 | 3.29 | .98 |
| 9* | Acctg | 25 | 3.72 | 1.14 | 18 | 3.22 | 1.17 |
| | Germ | 18 | 3.06 | 1.21 | 22 | 2.91 | 1.07 |
| | Math | 6 | 3.83 | .41 | 5 | 3.60 | 1.52 |
| | Soc | 76 | 3.66 | 1.07 | 38 | 3.00 | 1.07 |
| 10 | Acctg | 25 | 3.92 | .76 | 18 | 3.83 | .71 |
| | Germ | 18 | 4.17 | .79 | 22 | 3.82 | .96 |
| | Math | 6 | 4.00 | .63 | 5 | 4.00 | .71 |
| | Soc | 76 | 3.67 | .90 | 38 | 3.62 | .82 |
| 11*+ | Acctg | 25 | 3.76 | .88 | 18 | 3.44 | 1.04 |
| | Germ | 18 | 4.17 | .71 | 22 | 3.68 | .72 |
| | Math | 6 | 2.50 | .84 | 5 | 2.40 | .89 |
| | Soc | 76 | 3.61 | .88 | 38 | 3.21 | .86 |

*difference between achievement groups was significant (.05)
+difference among courses was significant (.05)

The analysis is clearly based on relative status of the listed courses only. In a different subset of courses the status of a given course could be quite different. Also, the reader should note again that the number of students in mathematics classes is small; hence, the evaluations of students in that course must be regarded with reservation.

## Discussion

The two types of analyses reported here address distinctly different questions. The first analysis looked at faculty ratings in terms of how high achieving sections of courses rate their instructors in contrast with how low achieving sections rate their instructors. If teaching makes a difference, "good" teachers should advance achievement more than "poor" teachers. If teacher rating scales are assessing effectiveness of teaching, those teachers who advance their students the farthest should receive higher ratings than those teachers whose classes advance less far. This is the argument that supports the first analysis of this study.

The results on the total scores on the rating scales did indeed show high achieving sections of courses reporting stronger ratings for their instructors than was true of low achieving sections. This finding supports the claim of validity for the faculty ratings.

However, not all items in the scale contribute to that claim. The items that appear to discriminate between high and low achieving sections dealt with the student's knowing what was to be done, with the instructor's preparation, enthusiasm, clear explanations, and making the class interesting. The instructor's dealing with examinations, and being available to the students for consultation, were not regarded differently by high and low achieving sections.

The second analysis, although it addressed a slightly different question, produced quite similar results. The second analysis identified high achievers as those students who did somewhat better than they were predicted to do (based on a combination of talent variables), and low achievers were those students who did clearly worse than they were predicted to do. This analysis focused on individual student achievement, regardless of how their section as a whole achieved. The hypothesis in this analysis was that if the student relates to an instructor in such a way as to produce effective work on that student's part, that student will rate the faculty member higher than will the student whose interaction with that instructor has not produced effective work.

The analysis did in fact show that on total scale scores the high achievers provided ratings for their instructors above those provided by low achievers. On individual items within the scale (a less reliable observation) high achievers provided more substantial ratings than did low achievers on

how much they had learned, the difficulty of the course, the instructor's enthusiasm, preparation, ability to make the course interesting, and reports of student progress.

The above data again speak to the validity of the faculty rating procedure. Students who find their circumstances in a class sufficiently favorable to achieve beyond expectations should rate their instructors higher than students who do not find circumstances leading to achievement. These findings appear to be tied to the extent to which faculty relate to individual students.

This second analysis may also reflect another possible circumstance. Students who characteristically achieve above expectations, regardless of whom their instructor may be, may, by their view of the total academic scene, see their instructors as demonstrating more favorable qualities than do students who achieve below their apparent ability level. It also may mean that students majoring in the subject - and hence more likely to over-achieve - find the instructor more interesting, the class better organized, etc., than non-majors - who are more likely to under-achieve.

In any case, if the circumstances in a class lead a student into achieving better than he/she is predicted to do, that student may well be expected to rate the instructor higher than does the student who achieves less well than predicted to do. A valid faculty rating scale, then, should produce data that reflect this relationship, and indeed this is the way the scale performed. This finding may in one sense be regarded as additional evidence of the validity of the evaluation procedure.

It is further interesting to note that two items:

8. The exams cover the most important aspects of the course, and
10. My instructor is regularly available for consultation,

showed no significant difference between achievers and non-achievers under either analysis. This does not mean that these aspects of teaching are unimportant. It simply says that achievers and non-achievers regard their instructors about equally on these behaviors. To this end these items do not discriminate among the instructors in this study.

## Conclusions

The following conclusions appear to be borne out by the data collected in this study:

1. The evidence supports the hypothesis of validity for faculty rating scales.
2. Students in sections that achieve well tend to rate their instructors higher than students in sections that do poorly, across all courses. The differences in ratings are primarily associated with the instructor's ability to provide direction for the students and to "come across" to the students.
3. Within individual sections of a course students who "over-achieve" rate their instructors higher than students who "under-achieve."
4. Although some teacher behaviors appear to be associated with achievement across all courses, other behaviors appear to be related to achievement in a subset of the observed courses only.
5. Students rate the quality of instruction quite differently from course to course.

## References


Costin, Frank, Greenough, W.T., Menges, R.J. Student ratings of college teaching: reliability, validity, and usefulness. Rev. of Educ. Research, 1971, 41:511-535.

Cronbach, L.J. & Ferby, L. How we should measure "change" - or should we? Psych. Bull., 1970, 74:68-90.

Frey, P.W. Student instructional ratings and faculty performance. Science, 1973, 182:83-85.

Frey, P.W., Leonard, D.W., & Beatty, W.W. Student ratings of instruction: validation research. Am. Educ. Res. J., 1975, 12, 435-447.

Gessner, P.K. Evaluation of instruction. Science, 1973, 180:566-570.

Greenwood, G.E., Hazelton, A., Smith, A.B., & Wave, W.B. A study of the validity of four types of student ratings of college teaching assessed on a criterion of student achievement gains. Res. in Higher Ed., 1976, 5:171-178.

McKeachie, W.J. Student ratings of faculty. AAUP Bull., 1969, 55:439-444.

Menges, R.J. The new reporters: Students rate instruction. In C.R. Pace, ed., New directions in higher education: evaluating learning and teaching. San Francisco: Jossey-Bass, 1973, pp59-75.

Rodin, M., Frey, P.W., Gessner, P.K. Student evaluation. Science, 1975, 187:555-559.

Rodin, M. & Rodin, B. Student evaluation of teachers. Science, 1972, 177:1164-1166.

Sullivan, A.M. & Skanes, G.R. Validity of student evaluation of teaching and the characteristics of successful instructors. J. Educ. Psych., 1974, 66:584-590.

Turner, R.L. & Thompson, R.P. Relationships and residual learning. Paper at the annual meeting of the AERA, Chicago IL, 1974.